Reinforcement Learning based Visual Attention with Application to Face Detection

Ben Goodrich, Itamar Arel
Department of Electrical Engineering & Computer Science
University of Tennessee, Knoxville
bgoodric@utk.edu, itamar@utk.edu

Abstract

Visual attention is the cognitive process of directing our gaze on one aspect of the visual field while ignoring others. The mainstream approach to modeling focal visual attention involves identifying saliencies in the image and applying a search process to the salient regions. However, such inference schemes commonly fail to accurately capture perceptual attractors, require massive computational effort and, generally speaking, are not biologically plausible. This paper introduces a novel approach to the problem of visual search by framing it as an adaptive learning process. In particular, we devise an approximate optimal control framework, based on reinforcement learning, for actively searching a visual field. We apply the method to the problem of face detection and demonstrate that the technique is both accurate and scalable. Moreover, the foundations proposed here pave the way for extending the approach to other large-scale visual perception problems.

1. Introduction

Selective visual attention is the mechanism by which we can rapidly direct our gaze towards objects of interest in our visual environment [8][10][5]. From an evolutionary viewpoint, this rapid orienting capability is critical in allowing living systems to quickly become aware of possible preys, mates or predators in their cluttered visual world. It has become clear that attention guides where to look next based on both bottom-up (image-based) and top-down (task-dependent) cues[6]. As such, attention implements an information processing bottleneck, only allowing a small part of the incoming sensory information to reach short-term memory and visual awareness. That is, instead of attempting to fully process the massive sensory input in parallel, nature has devised a serial strategy to achieve near real-time performance despite limited computational capacity: Attention allows us to break down the problem of scene understanding into rapid series of computationally less demanding, localized visual analysis problems. With computer games, selective visual attention can allow for devices that direct attention to the player as needed, allowing for intelligent human-computer interaction.

The study of visual attention is relevant to a myriad of scenarios in which actions are based on visual information from the environment [7]. Efficient and reliable attentional selection is critical because various cues appear amidst a cluttered mosaic of other features, objects, and events. Attention mechanisms enable preferential processing of particular locations in the visual field or specific features of objects. To cope with the massive amounts of information which we are exposed to, the brain is equipped with a variety of attentional mechanisms. These serve two critical roles. First, attention can be used to select behaviorally relevant information and/or to ignore the irrelevant or interfering information. In other words, you are only aware of attended visual events. Second, attention can modulate or enhance this selected information according to the state and goals of the perceiver. With attention, the perceivers are more than passive receivers of information. They become active seekers and processors of information, able to interact intelligently with their environment.

In this paper, we introduce a novel visual attention paradigm that is framed as a learning process rather than an inference one. A reinforcement learning methodology is utilized in which state information is provided in the form of foveated vision, while the reward function motivating the learning process is defined as a reflection of the confidence the system has in recognizing various objects. Learning from experience, which is intrinsic to reinforcement learning, allows the agent to become highly proficient at finding objects of interest in large visual fields. As a result, an autonomous unsupervised framework is described, which scales to very large visual scenes and is generally invariant to the nature of the images processed.

The rest of the paper is structured as follows. Section II briefly reviews existing visual attention methods. Section

978-1-4673-1612-5/12/$31.00 ©2012 IEEE

III introduces the proposed learning-based approach to the visual search task. In Section IV simulation results are presented and discussed, while in Section V the conclusions are drawn.

2. Visual Search Methods

One method is to exhaustively search the image for a face. Viola and Jones demonstrated that an exhaustive search of an image using cascade classification techniques can detect objects belonging to a class, such as a face[16][17]. While this method was demonstrated to be fast and robust, and to work quite well in some situations, exhaustively scanning an image can potentially be both error prone, and computationally expensive. Biological systems certainly do not exhaustively scan a field of view for a target, but rather rely on other cues from the environment to intelligently decide where to focus their gaze.

Saliency maps are another common way of approaching the problem of visual search. A saliency map is a two dimensional transform of the image, where areas of content should correspond to more salient regions. The actual features that make a region salient can vary depending on the type of target that is being searched.

There is some evidence that biological systems have neuron activations that correspond to specific types of saliencies, in particular the Difference of Gaussians (DoG) filter, which can be used as a measure of the variation on a local region, appears to correspond to neuron activations in the retina [4]. Similarly, the responses of orientation-selective neurons are usually obtained through convolution by Gabor Wavelets which resemble biological impulse response functions [9].

While this evidence indicates that saliency maps may provide a biological foundation for an approach to visual search, an entirely bottom up saliency map approach seems incomplete for several reasons. First and foremost, saliency maps are inherently application specific. Consequently, a saliency map that is aimed at locating one type of feature in one type of environment may not work well when locating another type of feature in a different environment. Another reason is that saliency maps alone fail to provide a top down learning mechanism that is characteristic of biological systems. The saliency maps in use generally involve fixed filters that are not adaptive in the sense that they can identify different features. Finally, the entire image must be scanned in a winner take all approach when attempting to locate a target.

For these reasons, we argue that saliency maps provide only one piece of the puzzle. A top down reward driven process may provide the other piece, as there is evidence that visual attention is a reward driven process [6]. The method proposed here provides such a reward driven process by framing the problem as a reinforcement learning process. A similar scheme was described in [12], where a small focal region of the image was extracted for deriving a state component, and actions were defined as saccades. However, that work was restricted to tabular form and histogram representation was used as a state construct, suggesting that it will not scale to multiple environments. This work advances the paradigm of using reinforcement learning for scalable visual search problems by exploiting neural networks for function approximation in concert with a foveated image model. By rewarding the agent for finding the target (or region of interest) we devise a training scheme that allows the system to become proficient at saccading to the target of interest.

3. Learning based Control Framework for Visual Attention

3.1. Model-Free Reinforcement Learning

Reinforcement learning problems are typically modeled as Markov Decision Processes (MDPs) [15]. An MDP is defined as a (S, A, P, R) tuple, where S stands for the state space, A contains all the possible actions at each state, P is a probability transition function $S \times A \times S \rightarrow [0, 1]$ and R is the reward function $S \times A \rightarrow R$. Also, we define π as the decision policy that maps the state set to the action set: $\pi : S \rightarrow A$. Specifically, let us assume that the environment is a finite-state, discrete-time stochastic dynamic system. Let the state space S be $S = (s_1, s_2, ..., s_n)$ and, accordingly, action space A be $A = (a_1, a_2, ..., a_m)$. Suppose at episode k, the RL agent detects $s_k = s \in S$, the agent chooses an action $a_k = a \in A(s_k)$ according policy π in order to interact with its environment. Next, the environment transitions into a new state $s_{k+1} = s^{'} \in S$ with the probability $P_{ss'}(a)$ and provides the agent with a feedback reward denoted by $r_k(s, a)$. The process is then repeated. The goal for the RL agent is to maximize the expected discounted reward, or state-value, which is represented as

$$V^{\pi}(s) = E_{\pi}\left\{\sum_{k=0}^{\infty}\gamma^{k} r_k(s_k, \pi(s_k))|s_0 = s\right\} \qquad (1)$$

where $\gamma(0 \leq \gamma < 1)$ is the discount factor and $E_{\pi}\{\}$ denotes the expected return when starting in s and following policy π thereafter. The equation above can be rewritten as

$$V^{\pi}(s) = R(s, \pi(s)) + \gamma \sum_{s^{'} \in S} P_{ss'}(\pi(s))V^{\pi}(s^{'}) \qquad (2)$$

where $R(s, \pi(s)) = E\{r(s, \pi(s))\}$ is the mean value of the reward $r(s, \pi(s))$.

However, in many practical scenarios, as in our case, the transition probability $P_{ss'}(a)$ and the reward function $R(s, \pi(s))$ are unknown, which makes it hard to evaluate the policy π. Q-learning [15] is one of the most effective

and popular algorithms for learning from delayed rewards in absence of the transition probability and reward function. In Q-learning, policies and the value function are represented by a two-dimensional lookup table indexed by state-action pairs. Formally, for each state s and action a, we define the Q value under policy π to be:

$$Q^{\pi}(s,a) = R(s,a) + \gamma \sum_{s' \in S} P_{ss'}(a) V^{\pi}(s') \qquad (3)$$

as the expected discounted reward starting from s, taking the action a, and thereafter following policy π. In Q-learning, a learned action value function Q directly approximates the optimal value function through a process called value iteration. Correspondingly, the state-action value update rule is given by

$$Q_{k+1}(s,a) = \begin{cases} Q_k(s,a) + \alpha\delta_k & \text{if } s_k = s, a_k = a \\ Q_k(s,a) & \text{otherwise} \end{cases} \qquad (4)$$

where $\delta_k = r_k + \gamma \max_{a' \in A(s')} Q_k(s', a') - Q_k(s,a)$ is called the *temporal difference error*, and α the learning rate.

When addressing large state and/or action spaces, tabular methods are found inadequate. They do not scale in terms of both store capacity and processing speed, given that such schemes involve exhaustive sweep of the state-action space. Function approximation methods provide a practical solution for approximating the value function when large spaces are considered. In particular, multi-layer perceptron neural networks are commonly used as non-linear function approximators, effectively replacing that tabular form estimation of the value function $(Q_k(s,a))$ with a neural network [15]. The underlying assumption is that proximity in state representation maps to similarity in state-action values - an assumption that holds true in many real-world applications. The neural network can be trained in a regular supervised learning manner, by defining the temporal difference error as the network error guiding weight adaptation.

In the proposed architecture, a feedforward neural network is utilized for value estimation and, indirectly, action selection, as illustrated in Figure 2. The inputs to the network consist of visual information relative to the current focal region in concert with the action to be taken next. This action represents the direction (in radians) that the agent is to shift its focal point to during the next time step. The magnitude of the shift is fixed, while the angle defines the action taken. There are 32 actions assumed, quantizing the angular directions to steps of 11.25 degrees. During each time step, all 32 possible actions are sequentially provided as input to the network for which 32 state-action value estimates are being produced. We apply a soft-max action selection by choosing the greedy action (i.e. the one with the highest value function) with probability 0.90, while the rest of the time a randomly selected action (out of the 31 non-greedy actions) is selected. Such scheme guarantees that sufficient exploration takes place while the agent learns an adequate policy. The neural network is thus trained using backpropagation whereby the error on the value estimate is the temporal difference error.

3.2. Foveated Imaging

One striking feature of vision in most biological systems is that of foveated vision [14]. The latter facilitates the trade-off of obtaining high detail for the focal area (area at which gaze is directed), while retaining sufficient information about the peripheral view. Consequently, the detail diminishes with the distance from the focal region. We chose to employ a popular model for foveated vision known as log polar mapping [13]. Log polar mapping is a simple image transformation that captures high detail in the center, with the level of detail fading exponentially as the distance from the center increases. Assuming the focal point is at the origin, and x and y are given relative to the focal point, the log polar transformation can be described as a Cartesian to polar coordinate transformation with a radius that increases exponentially, such that

$$\rho = \ln \sqrt{x^2 + y^2}, \quad \theta = \arctan\left(\frac{y}{x}\right), \qquad (5)$$

where ρ denotes the log radius and θ the angle to the origin.

3.3. Problem Formulation

In order to model visual search as a reinforcement learning problem, it is necessary to define the states, actions, and rewards considered. It would seem natural that the foveated image originating from a single fixation point would constitute a state. The actions would then be defined as movements away from the current fixation point (i.e. saccading in some direction). The reward construct can be defined in one of several ways. A scheme that offers reward for finding the target would seem the most natural. Ideally, under this framework, our agent could learn to find targets of a particular type across multiple images.

The state signal was defined as a function of log polar mapping with a radius of 128 pixels. Such mapping produced a 32x32 log polar image, yielding an input dimensionality that would have been too high for a single neural network to process. Thus, in order to reduce the dimensionality of the state space, we applied principal component analysis (PCA) to each input. To accurately determine the PCA matrix that would capture the most variance, it is necessary to first consider a set of typical observations. We chose to simply select a large random set of log polar images, across all training images, as the set of observations used in forming the PCA matrix. The goal of such selection is to capture as many variations as possible given the memory and run time requirements of the testing setup. This resulted in 32,768 random observations used. Once the PCA

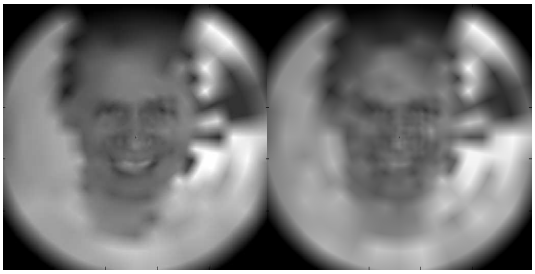

Figure 1. Illustration of log polar transform applied to a fovea image. On the right is the original 32x32 image and on the left is the reconstructed image, after PCA reduction was applied.

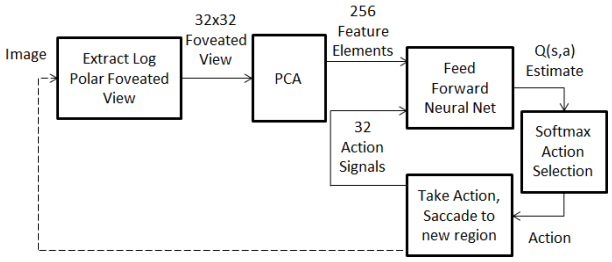

Figure 2. Data flow diagram for the proposed reinforcement learning based visual attention mechanism.

matrix was determined, it remained unchanged across the training and testing phases.

It was necessary to experiment with PCA to determine the appropriate number of principal components needed without losing too much information. Figure 1 depicts an example of a log polar image before and after PCA was applied. In order to generate these figures, we applied both inverse PCA and inverse log polar mapping. A criteria used for determining the appropriate number of principal components was retention of more than 95% of the signal variance. It was found that selecting the first 256 principal components met this criteria and resulted in recognizable reconstructed images, without loss of too much detail.

The action set represented 32 possible equidistant directions in which the gaze can be advanced. The distance of each saccade was fixed at 30 pixels. The image imposed a hard border preventing the agent from taking actions that would shift the gaze outside the image boundaries. Initially, a positive reward for finding the goal state and a reward of 0 otherwise was imposed. However, it was found that in early episodes the agent would often gravitate to the side of each image, or to one of its corners, and remain is these regions for prolonged periods of time. This was due, primarily, to a lack of direction or encouragement to adequately explore the image. By applying a small negative reward for attempting to move off the edge of the image, the agent was encouraged to remain within the bounds of the visual field until it identified the goal region, thus greatly reducing the duration of the initial episodes. Consequently, the revised reward function was given by

$$r(t) = \begin{cases} -1 & \text{when reaching image border} \\ 10 & \text{when finding the goal} \\ 0 & \text{else} \end{cases}$$

The goal region is defined by a circle with a radius of 20 pixels, the center of which is located between the eyes of each face. Figure 2 depicts the overarching data process flow in the proposed system. At its core, the system involves a feedback loop where by actions selected impact subsequent images processed. It is noted that there are no pre-determined features or image primitives provided to the system. As a result, the agent indirectly learns to capture regularities in face patterns in a completely autonomous manner.

4. Experimental Results

A face detection task was chosen to evaluated the proposed visual search process. All experiments performed referred to the BioID database [1], which consists of 1,521 images of human faces. Images are grey-scaled with a resolution of 384×286 pixels. All images are annotated with eye positions facilitating easy configuration of the goal region. The data set was partitioned into 1,000 images for training and 521 images for testing and experiments consisted of first training the system (i.e. adapting the weights of the neural network) to produce value function estimates using Q-Learning, as described above. The standard Q-Learning update rule was employed. The feed-forward neural network contained of a single hidden layer of 128 neurons, and an initial step size of .002. Stochastic Meta-Descent (SMD) [3][11] was employed as an adaptive step size scheme for the gradient descent updates. Input to the neural network consisted of a concatenation of 256 state/observation signals and 32 action signals, for a total of 288 inputs. The action signal was encoded such that all inputs were set to -1 with the exception of the action taken, which was set to $+1$ (i.e. one-hot encoding). The network output was a single scalar representing the state-action value estimate.

Once the agent identified the goal region, the appropriate reward was issued and the episode was terminated. New episodes were initiated at random locations of the image. Figure 3 provides a typical representation of how the training process evolved by depicting the amount of steps taken to reach the target as a function of the first 100,000 episodes. To ensure the network reached an adequate level of proficiency, the training process was executed until it reached a sufficient steady state such that episode durations were all below 20 steps. Once the training process exceeded 150,000 episodes, the trained network was loaded into the testing environment. The test runs were set up in a similar manner

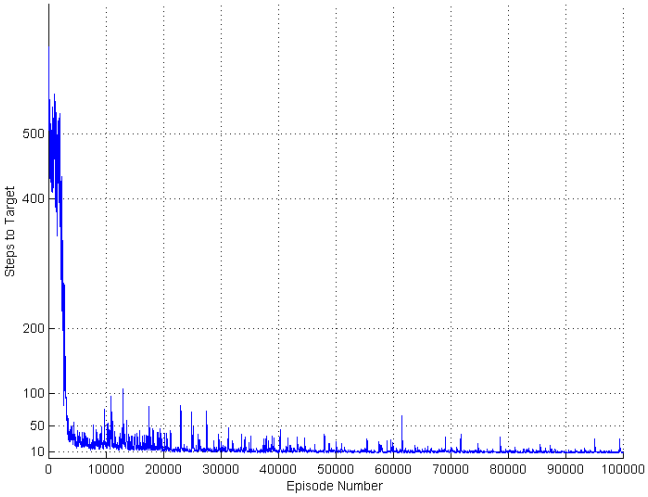

Figure 3. Progression of the episode duration as a function of the episodes during the training process.

to that of the training runs, with the exception that weights were kept unchanged

For comparison to an existing method, the widely popular Viola Jones face detection method was used[16][17]. OpenCV's built in traincascade utility was given a similar training set taken from the same 1,000 training images [2]. The positive examples were created from the eye positions by creating bounding boxes from the eye position data. The bounding boxes were created such that their width and height was 150% of the eye width, providing a consistent view of the upper half of the face and allowing the detector to be trained to find the position in between the eyes such that it could be compared with our detector.

Q-learning is essentially a stochastic process that depends on the number of steps we chose to run, as well as the starting position. To compare our method consistent results were needed. To achieve this, Episodes were initiated multiple times at random locations on the same image and ran for a fixed number (i.e. 50) of steps. The final point after running 50 steps was recorded. An average was taken of all of the final points after approximately 150 runs per image to remove any noise in the process. The position of this average final point was compared to the position in between the eyes where the detector was trained to locate.

Table 1 summarizes the results. For the Viola Jones face detector, the center of the rectangle was used. If the detector found more than one rectangle in an image, only the largest one was used. This seemed reasonable given that in every image that was inspected with multiple rectangles, the largest one corresponded to the face. If the center of the rectangle fell within a certain pixel distance (first column in the table) it was considered a successful detection. Our method was evaluated in a similar manner. If the average final point over multiple episode runs was found to

Pixel Distance From Center	Q Learning Detection Rate	Viola Jones Detection Rate
10	40.11%	89.64%
20	79.85%	90.02%
30	92.89%	90.02%
40	97.69%	90.02%
50	99.62%	90.02%

Table 1. Success rates of the Viola Jones face detector algorithm, compared to those of the proposed approach, as a function of various radii (measured in pixels) from a focal area of each face.

Figure 4. Illustration of different pixel distance metrics used for determining success/failure. The circles from smallest to largest represent a radius of 10, 20, 30, 40, and 50 pixels, respective

be within a predefined pixel distance, it was considered a successful detection. The table summarizes the percentage of successes out of the 521 test images using various pixel distance metrics.

4 illustrates what different distance metrics correspond to in terms of one of the sample image. Five circles have been depicted around the center of the subject's eyes, with a radii of 10, 20, 30, 40, and 50 pixels, respectively. Note that the overall scale of the face does vary somewhat in the BioID dataset, and this is only one example. As can be observed from these results, the proposed method generally converges to the region of the face, suggesting that the network is accurately learning a generalized behavior for locating faces. It's important to note that optimality is not a direct goal of the proposed method. While saccading along an optimal trajectory on route to finding a target would be a desired goal, it is not perceived as a pragmatic one. It can be argued that even mammals do not exhibit optimal visual search in every setting. However, a consistently well-performing, sub-optimal visual search engine that exhibits solid performance across different image domains, is a practical and notable achievement. A realistic aim for the visual attention mechanism described would be to yield search trajectories that are no more than 30% longer in duration to

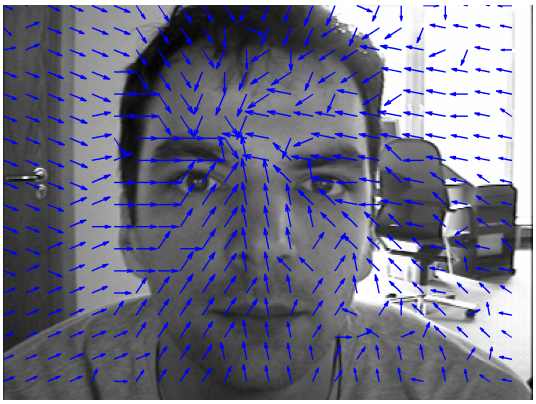

Figure 5. Illustration of the agent's preferred direction of saccading for uniformly sampled locations across the image. Direction preference is determined by the state-action value function estimate.

that of the optimal ones.

Figure 5 illustrates the direction that the agent suggested taking at uniformly sampled locations in an image taken from the test set. The suggested direction to shift the gaze was obtained directly from the state-action value estimates. As can be observed, the actions direct the gaze to the center of the face, which is the desired outcome.

5. Conclusions

This paper introduced a novel approach to modeling visual attention processes - one that is based on framing the problem as a reinforcement learning task. The system hosts a feedfoward neural network for state-action value estimation in concert with a fovea vision model. The latter is a biologically plausible construct that resulted in modest state representation. The proposed framework has several key advantages relative to existing schemes. One advantage is that it does not require exhaustively processing the entire image, rendering it highly scalable. Another key advantage is that our method avoids the need for manually define saliencies and hand-crafting visual features, as is commonly required by existing saliency-based schemes. The framework does not preclude including saliency information as additional features, although that was not explored in this paper. The method introduced merely provides a reward-driven, top down framework that was successfully applied to a face detection task, illustrating its potential in serving as a generic engine for active visual search. Moreover, the approach naturally scales to large visual fields and can be applied to other challenging machine vision tasks, such as object tracking and segmentation.

References

[1] BioID Face Database. http://support.bioid.com/downloads/facedb/index.php. [Online; accessed 21-Nov-2011]. 4

[2] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 4

[3] M. Bray, E. Koller-Meier, P. Müller, L. V. Gool, and N. N. Schraudolph. 3D Hand Tracking by Rapid Stochastic Gradient Descent Using a Skinning Model. In *First European Conference on Visual Media Production (CVMP)*, pages 59–68, London, 2004. 4

[4] P. Cagnac, N. Di Noia, C.-H. Huang, D. Racoceanu, and L. Chaudron. Consciousness-driven model for visual attention. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 1061 –1066, 31 2011-aug. 5 2011. 2

[5] L. Chelazzi, E. Miller, J. Duncan, and R. Desimone. A neural basis for visual search in inferior temporal cortex. *Nature*, 363:245–247, 1993. 1

[6] J. Chun, M.M. & Wolfe. Visual attention. *Blackwell handbook of perception*, pages 272–310, 2001. 1, 2

[7] A. F. C. B. Emanuela Bricolo, Tiziana Gianesini and L. Chelazzi. Serial attention mechanisms in visual search: A direct behavioral demonstration. In *J. Cognitive Neuroscience*, pages 14:980–993, October 2002. 1

[8] L. Itti and editor J. Feng. Modeling primate visual attention. In *Computational Neuroscience: A Comprehensive Approach*, pages 635–655. CRC Press, Boca Raton, 2003. 1

[9] L. Itti and C. Koch. Computational modeling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001. 2

[10] T. Koide and J. Saiki. Stochastic guided search model for search asymmetrics in visual search tasks. In H. Bultho, editor, *Lecture Notes in Computer Science*, volume 2525/2022, pages 67–111. 2002. 1

[11] Z. Liu and I. Elhanany. A fast and scalable recurrent neural network based on stochastic meta descent. *IEEE Transactions on Neural Networks*, 19(9):1652–1658, 2008. 4

[12] S. Minut and S. Mahadevan. A reinforcement learning model of selective visual attention. In *In Proceedings of the Fifth International Conference on Autonomous Agents*, pages 457–464. Press, 2000. 2

[13] R. Peters II, M. Bishay, and T. Rogers. On the computation of the log-polar transform. *image*, 1:1, 1996. 3.2

[14] N. M. Putnam, D. X. Hammer, Y. Zhang, D. Merino, and A. Roorda. Modeling the foveal cone mosaic imaged with adaptive optics scanning laser ophthalmoscopy. *Opt. Express*, 18(24):24902–24916, Nov 2010. 3.2

[15] R. S. Sutton and A. G. Barto. In *Reinforcement Learning: An Introduction*. MIT Press, 1998. 3.1, 3.1, 3.1

[16] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001. 2, 4

[17] P. Viola and M. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004. 2, 4